SCHRAUSSER

Randomisierungstests: Eine Einführung in die Method

Institut für Psychologie, Universität Graz, Universitätsplatz 2 8010 Graz, Austria

Allgemeines

Permutationstests wurden das erste Mal von Fisher (1935) behandelt. Diese Tests liefern dann valide Signifikanztests, wenn man keine Zufallsstichprobe aus einer definierten Population zur Verfügung hat. Die Beobachtungseinheiten oder Versuchspersonen müssen per Zufall den verschiedenen experimentellen Bedingungen zugeteilt werden. So bezieht sich der Zufall nicht auf die Stichprobenauswahl, sondern auf die Zuteilung der Personen. Fisher bezeichnete das Verfahren allerdings als unpraktisch wegen des erhöhten Rechenaufwandes und der 1935 noch nicht sehr ausgereiften Rechenmaschinen.

Pitman und Welch (1937) überprüften nur die ersten 4 Momente einer Permutationsverteilung der Teststatistik (AM, Varianz, Schiefe, Exzeß) und konnten große Übereinstimmungen mit bekannten Verteilungen (NVT) feststellen. In den 50er und 60er Jahren wurden eine Reihe von mathematischen und theoretischen Arbeiten zu diesem Thema veröffentlicht, welche die Effizienz von Permutationstests bestätigen.

Dwass (1957) schlägt vor nicht die exakte Permutationsverteilung der Teststatistik auszunutzen, das heißt nicht alle möglichen Permutationen durchzuführen, sondern nur eine Zufallsstichprobe (mit oder ohne Zurücklegen) aus allen möglichen Permutationen zu verwenden. Diese Vorgangsweise wird allgemein als Monte-Carlo Version des Permutationstests (MC) bezeichnet, oder auch als randomisierte Permutation (Edgington, 1980). Auf diese Weise, so Willmes (1987) sind Permutationstests wirklich in der Praxis einsetzbar, da der Rechenaufwand erheblich gesenkt werden kann. Bei Untersuchungen konnte gezeigt werden, daß die Einbuße an Teststärke gegenüber der exakten Permutation nicht gravierend ist. Nicht zuletzt ermöglichen die steigenden Rechnergeschwindigkeiten einen unaufwendigen Einsatz von MC Permutationstests in der Praxis. Eine größere Anzahl von Computerprogrammen zu MC Tests wurden publiziert darunter sind Arbeiten von Green (1977), Hawel (1982), Heller (1981), Pagano & Tritchler (1983), Still & White (1981) und Streitberg & Röhmel (1986). Willmes stellt weiters fest, daß die Anzahl der Publikationen die Permutationstests behandeln ansteigt.

Da bei Permutationstests immer nur mit einer spezifischen Verteilung, die nur für dieses eine Experiment Gültigkeit besitzt und deshalb nicht generalisierbar ist gearbeitet wird, ist auch die Bestimmung der Teststärke *vor* der Untersuchung nicht möglich. Willmes spricht hier auch von der bedingten Teststärke des Permutationstests.

Auch in der Bezeichnung der Methode herrschen Unklarheiten. Es wird vorgeschlagen den Namen ‘Permutationstests’ dann zu verwenden, wenn Zufallsstichproben aus unbekannten Populationsverteilungen verglichen werden. Der Terminus

Randomisierungstests soll dann verwendet werden, wenn die Ergebnisse von Beobachtungseinheiten (Versuchspersonen) verglichen werden, die aufgrund experimenteller Randomisierung, das heißt durch zufällige Zuteilung zu den experimentellen Bedingungen zustande kamen.

Willmes (1987, S.8 ff) bespricht auch die multivariate Variante des Permutationstests, auf welche hier nicht näher eingegangen wird. Die Bedeutung dieser Verfahren liegt u.a. darin, daß die Korrelationen zwischen den abhängigen Variablen mitberücksichtigt werden. Allerdings ist multidimensionale Normalverteilung, eine Grundvoraussetzung zur Anwendung von multivariaten Verfahren besonders in der Psychologie so gut wie nie gegeben. Es wird festgestellt, daß Permutationstests für multivariate Designs noch in den Kinderschuhen stecken.

Der erste Übersichtsaufsatz über die Verwendung von Permutationstests in der Psychologie stammt von Moses (1952). Weiters werden die Publikationen von Edgington (1964, 1966, 1969, 1969a, 1969b, 1975, 1980, 1987, 1995) lobend erwähnt. Edgington betont die Validität des Randomisierungstests und streicht die Bedeutung der Methode heraus, wenn man keine Zufallsstichproben verwenden kann, so wie es in der psychologischen Forschung nur allzu oft der Fall ist. Er sagt, daß einzig und alleine eine randomisierte Zuweisung der Beobachtungseinheiten zu den experimentellen Bedingungen prinzipiell und alleine wichtig ist. Die Generalisierung ist kein statistisches Verfahren, sondern ein induktiver Schluß. Wie man den Schluß vom speziellen auf den allgemeinen Fall rechtfertigt, ist kein statistisches Problem mehr. Willmes (1987) kritisiert Edgington (1980, S.16) „Parametric tests of all kinds, including relatively complex tests like factorial analysis of variance [...] are distribution-free tests when the significance of the teststatistic is determined by permuting the data“, zumal dieser später selbst bemerkt, daß es keinen validen Randomisierungstest auf Wechselwirkungen in mehrfaktoriellen Plänen gibt, “Even for the very simplest of factorial designs for independent groups [...] there is no way to test the H_0 of no interaction by data permution“ (S.162). Auch Phyel (1978, 1980) hat skizziert, daß es keinen invarianten Permutationstest auf Wechselwirkungen gibt.

Edgington (1980) berichtet lobenswerterweise sehr viel über Einzelfallexperimente, worauf Willmes nicht eingeht. Weiterführende Literatur zu Einzelfallexperimenten findet der interessierte Leser bei Petermann & Hehl (1979).

Zum Stellenwert der Permutationstests in einer deduktivistischen Theorie des psychologischen Modells wird auf Bredenkamp (1980) verwiesen. Dieser sagt, daß es in einer deduktivistischen Theorie möglich sein muß auch eine Entscheidung über die Gültigkeit der Nullhypothese zu treffen. Mit Permutationstests ist es jedoch nicht möglich vor dem Experiment den für eine gewünschte Teststärke erforderlichen Mindeststichprobenumfang festzulegen.

Fragestellungen sind noch, für welche Tests man analoge Invarianz-Permutationstests konstruieren kann Trendtests, oneway, Wechselwirkungen, ANCOVA, Multiple-Regression, Verlaufskurven unter verschiedenen experimentellen Bedingungen, Überlebensanalysen und besonders Einzelfallstudien (siehe Edgington 1975, 1980) und vor Allem Pitman (1937a, 1937b, 1938).

Prinzip des Permutationstests (Willmes 1987, S.26 ff)

Die konzeptuellen Unterschiede zwischen Permutationstest und Randomisierungstest, auf welche sich Willmes (1987) bezieht wurden von Kempthorne (1976) ausgearbeitet.

Permutationstest:

Beim Permutationstest werden Stichproben aus einer Grundgesamtheit analysiert oder verglichen. Hier ist wichtig, daß es sich um Zufallsstichproben aus der Population handelt. Die Verteilungen in Population und Stichprobe muß nicht näher bekannt sein. Es werden solche Hypothesen getestet, die annehmen, daß, die gemeinsame Verteilung der Werte invariant gegenüber jeder Permutation bleibt. Die Teststatistiken der Permutationen werden nach der Größe geordnet aufgetragen. So erhält man die empirische Permutationsverteilung der Teststatistikwerte in einer Treppenfunktion. Unter H_0 ist die beobachtete Verteilung der Werte gleich der Permutierten Verteilungen. Die Signifikanz wird ermittelt als die Wahrscheinlichkeit, daß man Testwerte erhält, die größer oder gleich dem beobachteten Testwert sind. P wird als die Konsistenz der beobachteten Daten mit der Nullhypothese interpretiert, je kleiner *p*, desto größer ist die Evidenz für die Alternativhypothese. Der Test kann immer nur ein Signifikanzniveau annehmen, das ein ganzzahliges Vielfaches von $1/n(\Pi)$ (1 durch Anzahl der Permutationen) ist.

Randomisierter Permutationstest:

Hier werden nur eine gewisse Anzahl der Teststatistikwerte, die gleich dem beobachteten Teststatistikwert sind zur Berechnung des Signifikanzniveaus verwendet. Es wird ein Anteil von 50% $T^{=}$ empfohlen, 'mid-*p*' (Franck, 1986). So kann die Teststärke erhöht werden. Der randomisierte *p*-Wert ist eher progressiv, der nicht-randomisierte eher konservativ.

Permutationstests sind sogenannte bedingte Tests, das heißt die Verteilung der Variablen ist abhängig von der Auswahl der Variablen, die in jedem Experiment anders ist. Aus diesem Grunde muß in jedem Experiment eine *p*-Wert Bestimmung durch Permutation erfolgen.

Permutationstests im Randomisierungsmodell, Randomisierungstest oder Re-Randomisationstests:

Hier ist eine Zufallsauswahl einer Stichprobe aus einer Population nicht notwendig, es ist nur wichtig, daß die Personen per Zufall den Bedingungen eines Versuchsplanes zugeteilt werden (vgl. Bredenkamp, 1980). Als Grundgesamtheit oder Population ergibt sich also die Gesamtheit aller Möglichen Versuchspläne oder Zuordnungen. Unter der Nullhypothese haben alle Versuchsbedingungen für alle Personen den gleichen Effekt. Eine Variation zwischen den Personen kann sich nur als eine Variation innerhalb der Personen ergeben, das ist die Fehlervarianz der Versuchspersonen (Einfluß der 'unit-values'), wenn Meßfehler nicht berücksichtigt werden. Die Teststatistik ist eine Funktion der Variabilität der Personen ('unit-values') und der Variabilität der Faktoren, oder der Variabilität durch die Zuteilung zu den Faktoren ('Designervariable').

Randomisierter Randomisationstest:

Die Definition ist analog zum randomisierten Permutationstest. Da die Verteilung der Werte abhängig von den jeweiligen unit-values ist, die in jedem Experiment anders ist, ist eine Vertäfelung der kritischen *p*-Werte, so wie beim Permutationstest nicht möglich.

Der Unterschied zwischen Permutations- und Randomisationstest ist also der, daß sich das Zufallsmoment beim Permutationstest auf die Auswahl der Personen und beim Randomisationstest auf die Zuteilung von Personen bezieht. Während die Grundgesamtheit des Permutationstests die Population, aus der die Stichprobe gezogen wurde, ist die

Grundgesamtheit des Randomisationstests die Gesamtzahl aller möglichen Zuteilungen zu den Versuchsbedingungen oder anders ausgedrückt alle möglichen Versuchspläne. Der Ansatz von Pyhel (1980) zum Vorgehen bei der Konstruktion von Invarianz-Permutationstests fordert, daß die Verteilungsfunktion der beobachteten Werte gleich ist wie die Verteilungsfunktion aller Permutierten Werte. Wichtige Versuchspläne, für die Invarianz-Permutationstests für Haupteffekte existieren sind:

- *Das unverbundene 2-Stichprobenproblem (Unabhängiger t-Test)*
- *Das verbundene 2-Stichprobenproblem (Abhängiger t-Test)*
- *Das unverbundene k-Stichprobenprobem oder der einfaktorielle Zufallsplan*
- *Das verbundene k-Stichprobenproblem oder der einfaktorielle Blockplan*
- *2 und 3-faktorielle Versuchspläne*
- *Das lateinische Quadrat*

Für Wechselwirkungseffekte existieren bis dato [1987] noch keine mathematischen Herleitungen. Die Teststatistiken sind, sofern diese in einer monotonen Beziehung zueinander stehen frei wählbar (vgl. Edgington, 1980).

Die einzige schwache Verteilungsannahme bei Randomisationstests ist die Permutationsinvarianz der Fehlerterme (‘unit-values’), was durch zufällige Zuteilung der Personen zu den Versuchsbedingungen erreicht wird (vgl. Bredenkamp 1980, S.18). Das zweite Fehler- oder Zufallselement nach den unit-values stellt der Meßfehler oder ‘technical-error’ dar, der unabhängig von der randomisierten Zuteilung der Personen ist.

Teststärke(S.245ff):

Ein gutes statistische Verfahren soll einen vorhandenen Unterschied in der Population auch erkennen können, d.h. möglichst teststark sein. Die Teststärke, die sich in parametrischen Verfahren prinzipiell bestimmen läßt, ist in Permutations oder Randomisationstests vor Ausführung des Experiments nicht bestimmbar, da sich die Verteilung der Variablen von Experiment zu Experiment unterscheidet, und die Verteilung in der Grundgesamtheit bekannt sein muß. Somit kann auch der erforderliche Mindeststichprobenumfang zur Aufdeckung eines Effektes nicht bestimmt werden. Die Teststärke kann also immer nur konditional zu den aktuellen unit-values a posteriori bestimmt werden. Edgington (1980) bezeichnet die Teststärkenproblematik als ‘second-order problem’ und befaßt sich gar nicht mit diesem Thema. Es hat sich gezeigt, daß der Randomisationstest in ANOVA-Modellen die gleiche Teststärke besitzt wie parametrische Verfahren. Approximative Methoden zur Bestimmung der bedingten a-posteriori Teststärkebestimmung werden besprochen [hier sollte man sich etwas einarbeiten] (S. 252ff). Willmes (1987) empfiehlt, daß die a-posteriori Teststärke zusätzlich zu den Ergebnissen angegeben wird.

Robustheit (S.286ff):

Ein statistisches Verfahren wird allgemein als ‘robust’ bezeichnet, wenn es wenig empfindlich gegenüber Verletzungen der Voraussetzungen ist. Für den Permutationstest ist das Konzept der qualitativen Robustheit von Interesse. Die qualitative Robustheit eines Tests bedeutet, daß mit kleinen Änderungen in den Personen, oder der Personenverteilung mit kleinen Änderungen des p-Wertes ($-1/n \log p_s(x)$) einhergehen. Es konnte bewiesen werden, daß diese Annahme weder für den t-Test noch für den Permutationstest der Fall ist. Er schlägt deshalb eine robustere Teststatistik als den t-Test für den Permutations- oder auch

Randomisationstest vor. Die Bedeutung der Robustheit ist bis dato [1987] als zweitrangiges Problem behandelt worden, auch Willmes (1987) hat dazu keine Untersuchungen angestellt.

Die Bedeutung der Permutationstests oder randomisierten Permutationstests in der Psychologie

Willmes (1987) bezieht sich neuerlich auf Bredenkamp (1980) und dessen Ausführungen über die deduktivistische Theorie des Experiments, das ohne induktive Verallgemeinerung auskommen soll. Für die Validität eines Verfahrens ist es wichtig, daß es sich um Zufallsstichproben handelt. Solche echten Zufallsstichproben können aber nur aus existierenden endlichen Populationen gezogen werden, was in psychologischen Untersuchungen kaum der Fall ist. Mit Randomisierungstests, deren zufällig zugeteilte Beobachtungseinheiten eine endliche Grundgesamtheit sind, kann man hingegen valide teststatistische Nullhypothesen prüfen. Diese sind per Konstruktionem verteilungsfreie (= Populationsverteilung ist unwichtig), bedingte (= gelten nur für das eine Experiment), exakte und finite (= endliche Grundgesamtheit) Tests zum Niveau α. Eine weitere Voraussetzung ist die Additivität der Effekte der 'unit-values' und der 'Designervariable'. Auch wenn man nicht deduktivistisch, sondern induktivistisch argumentiert sind Randomisierungstests als Erkundungsexperimente ein gutes Hilfsmittel (vgl. Bredenkamp, 1980), wenn man am Beginn eines Forschungsprojektes noch keine Verteilungsannahmen treffen kann. Der induktive Schluß von den untersuchten auf andere Beobachtungseinheiten der selben Population ist nicht statistisch sondern wissenschaftlich, 'scientific-inference' (vgl. Edgington, 1980, S.5; Kemthorne, 1952, S.152; Welch, 1937, S.47). Die Bestätigung einer Hypothese ist jedoch unmöglich. Das Resultat des Permutations- oder Randomisationstests kann nicht auf die Population verallgemeinert werden.

Praktische Ausführung von Permutations- und Randomisierungstests

Bradley (1968) kritisiert das Verfahren als nicht praktikabel und undurchführbar. Auch Bredenkamp (1980, S.20) spricht von unerhörtem Rechenaufwand dieser Verfahren. Folgende Auswege aus diesem Dilemma werden beschrieben:

A*pproximationen:*

Für das 1 2 und *k* Stichprobenproblem wurden Approximationen zur exakten Permutationsverteilung vorgeschlagen (Reihenentwicklungen). Gute Übereinstimmung bestehen im 3 und 4 Moment der Verteilungen des exakten und approximativen *p*-Wertes.

Partialsummen:

Durch den Vergleich von Partialsummen aus den Beobachtungen mit dem Teststatistikwert kann 90% der möglichen Summenbildung entfallen.

Verschiebealgorithmus, 'shift-algorithmus' (Streitberg & Röhmel, 1983, 1984, 1986):

Mit wachsendem *n* besteht nur ein polynomialer Anstieg der Rechenoperationen. Allerdings dürfen nur positive, ganzzahlige Werte verwendet werden und der Bedarf an Speicherplatz ist erheblich. Willmes (1987) liefert einige FORTRAN77 Programme, die mit dem Shift-Algorithmus arbeiten. Für komplexere Versuchspläne als das (un)verbundene 2 oder *k* Stichprobenproblem im uni und multivariaten Fall sind noch keine Shift Algorithmen bekannt.

Sukzessive systematische Permutation durch schnelle Algorithmen:

Einige Autoren geben einen guten Überblick über FORTRAN77 Algorithmen für das *k*-Stichprobenproblem. Beispielsweise Permutation durch Inversion.

Testentscheidung zugunsten einer Hypothese:

Will man nicht den exakten *p*-Wert bestimmen, sondern nur zugunsten einer Hypothese bei gegebenem Signifikanzniveau und gegebener Anzahl von Permutationen entscheiden, so kann man den Vorgang von Permutation, Berechnung der Teststatistik und Vergleich so lange durchführen, bis man die H_0 das erste Mal annehmen kann. Es wird also überprüft, ob die Anzahl der extremen Teststatistikwerte größer oder gleich dem αten Teil auszuführender Permutationen ist (vgl. Jöckel, 1982, S.72ff).

Monte-Carlo Version:

Verwendung einer (großen) Zufallsstichprobe aller möglichen Permutationen (Dwass, 1957). Dies Vorgangsweise wird als Monte-Carlo Version des Permutations- bzw. Randomisationstest bezeichnet. Um 1933 wurde erstmals das Prinzip eines Monte-Carlo Tests zur Überprüfung der Robustheit einer Teststatistik gegenüber der Verletzung der Normalverteilung. Edgington (1969b) bezeichnet die MC-Version des Randomisationstest als approximativ, was laut Willmes (1987, S.393) nicht korrekt ist, da auch der MC-Randomisationstest ein exakter Test zum Niveau α bleibt. Auch die Einbuße an Teststärke bei einer Anzahl von über 10000 Permutationen ist sehr gering ist. Die Teststärke der MC-Version kann bei wachsendem Stichprobenumfang nicht abnehmen (Jöckel, 1982).

Im Jahr 1966 wurde ein MC Permutationsprogramm für *F*-Test Statistiken publiziert. Jöckel (1982) warnt vor zu kleinen Umfängen an Permutationen im Bereich zwischen 500 und 1000, wie sie auch sehr häufig von Edgington verwendet werden. Willmes (1987, S.399) zeigt eine Tabelle mit Konfidenzintervallen von *p*-Werten bei gewissen Stichprobengrößen von Permutationen. Die sogenannte Dwass Effizienz (Willmes, 1987, S.400) gibt an, wie groß die Mindeststichprobe an Permutationen sein muß, um einen vorgegeben Wert nicht zu unterschreiten oder wie groß die Abweichung zur Teststärke aus einer systematischen Permutationsverteilung ist. Diese Dwass Effizienz oder auch die Anzahl von Permutationen sollte, im Gegensatz zur Meinung Edgingtons (1980) im Computerausdruck mit angegeben werden.

Ein weiterer Punkt ist die Problematik der Bildung von exakt gleichverteilten Zufallszahlen, aufgrund derer eine Permutation ausgewählt wird. Ist die Bildung von Zyklen in den Zufallszahlen zu rasch, so wird die Voraussetzung der Gleichwahrscheinlichkeit der Auswahl jeder Permutation verletzt. Da die Stichproben mit zurücklegen gezogen werden, was einen erheblich geringeren Rechenaufwand mit sich bringt, ist es auch in dieser Beziehung wichtig gleichverteilte Zufallszahlen zu errechnen, da dadurch die Wahrscheinlichkeit einer wiederholten Ziehung sinkt. Es konnte nachgewiesen werden, daß zwischen Ziehungen mit und ohne Zurücklegen keine Unterschiede im *p*-Wert auftraten.

Randomisierter MC Randomisationstest, [Willmes würde sagen 'Randomisierter Monte-Carlo Permutationstest im Randomisierungsmodell oder Randomisierter Monte-Carlo Re-Randomisationstest']:

Hier wird wiederum eine gewisse Anzahl (50%, 'mid-*p*') von $T^{=}$ In die *p*-Wert Berechnung mit aufgenommen. Im Gegensatz zum Nicht-randomisierten MC Randomisationstest oder Nicht randomisierten Monte Carlo Permutationstest im

Randomisierungsmodell = Nicht randomisierter Monte Carlo Re-Randomisationstest oder auch schlicht Monte Carlo Randomisationstest ohne Randomisierung = Monte Carlo Permutationstest im Randomisierungsmodell oder wie Edgington (1980) sagen würde der randomized randomized permutation test, der alle $T^{=}$ in die Berechnung des *p*-Wertes aufnimmt.

Literatur

Bradley, J. V. (1968). *Distribution-free statistical tests.* Englewood Cliffs: PrenticeHall.

Bredenkamp, J. (1980). *Theorie und Planung psychologischer Experimente*. Darmstadt: Steinkopff.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics, 28,* 181-187.

Edgington, E. S. (1964). Randomization tests. *Journal of Psychology, 57,* 445-449.

Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin, 66,* 485-487.

Edgington, E. S. (1969). Approximate randomization tests. *Journal of Psychology, 72,* 143-149.

Edgington, E. S. (1969a). *Statistical inference: The distribution free approach*. New York: McGraw Hill.

Edgington, E. S. (1969b). *Statistical inference: The distribution-free approach*. New York: McGraw-Hill.

Edgington, E. (1975). Randomization Tests for One-Subject Operant Experiments. *The Journal of Psychology, 90*, 57-68.

Edgington, E. S. (1980). *Randomization tests.* New York: Marcel Dekker.

Edgington, E. S. (1987). *Randomization tests (2nd ed).* New York: Marcel Dekker.

Edgington, E. S. (1995). *Randomization tests (3rd ed).* New York: Marcel Dekker.

Fisher, R. A. (1966). *The design of experiments (8th ed.)*. Edinburgh: Oliver & Boyd. (Originalausgabe 1935).

Franck, W. E. (1986). P-values for discrete test statistics. *Biometrical Journal, 28,* 403-406.

Green, B. F. (1977). Practical interactive program for randomization tests of location. *The American Statistican, 31*, 37-39.

Hawel, W. (1982). Algorithm and FORTRAN program for the permutation test with dependent samples. In H. Caussinus, P. Ettinger, & J.R. Mathieu (Eds.), *Compstats 1982: Part II (Supplement): Short communications, summaries of posters (p. 133).* Wien: Physica.

Heller, R. (1981b). Pitman's one-way analysis of variance: exact p-values computed by a PL/1 program. EDV in *Medizin und Biologie, 12,* 81-82.

Jöckel, K.-H. (1982). Eigenschaften und effektive Anwendung von Monte-CarloTests. *Dissertation,* Universität Dortmund, Abteilung Statistik, Dortmund.

Kempthorne, O. (1952). *The design and analysis of experiments*. New York: Wiley.

Kempthorne, O. (1976). Of what use are tests of significance and tests of hypothesis. *Communication in Statistics - Theory and Methods. A5,* 763 - 777.

Moses, L. E. (1952). Nonparametric statistics for psychological research. *Psychological Bulletin, 49*, 122-143.

Petermann, F., Hehl, F. J. (1979). Einzelfallanalyse-ein Überblick. In: Petermann, F., Hehl, F. J. (Hrsg.): *Einzelfallanalyse*. München: Urban & Schwarzenberg, S. 114.

Pagano, M. & Tritchler, D. (1983). On obtaining permutation distributions in polynomial time. *Journal of the American Statistical Association, 83*, 435-440.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. *Journal of the Royal Statistical Society Series B, 4*, 119-130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations II: The correlation coefficient test. *Journal of the Royal Statistical Society Series B, 4*, 225-232.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any populations. III. The analysis of variance test. *Biometrika, 29*, 322-335.

Pyhel, N. (1978). Beiträge zur Theorie der Permutationstests mit Anwendung in einem linearen Modell. *Dissertation*. Mathematisch-Naturwissenschaftliche Fakultät, RWTH Aachen, Aachen.

Pyhel, N. (1980). Distribution-free r sample tests for the hypothesis of parallelism of response profiles. *Biometrical Journal, 22*, 703-714.

Still, A.W. & White, A.P. (1981). The approximate randomization test as an alternative to the F-test in analysis of variance. *British Journal of Mathematical and Statistical Psychology, 34,* 243-252.

Streitberg, B., & Röhmel, J. (1983). Exakte Verteilungen für Rang- und Randomisierungstests im allgemeinen c-Stichprobenproblem. *Vortrag auf der Tagung der International Biometric Society, Section Deutschland*, März, Bad Nauheim.

Streitberg, B., & Röhmel, J. (1984). Exakte Verteilungen für Rang- und Randomisierungstests im allgemeinen c-Stichprobenproblem. *Unveröffentlichtes Manuskript*.

Streitberg, B., & Röhmel, J. (1986). Exact distribution for permutation and rank tests: An introduction to some recently published algorithms. *Statistical Software Newsletter, 12*, 10 - 17.

Welch, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika, 29*, 21-52.

Willmes, K. (1987). Beiträge zu Theorie und Anwendung von Permutationstests in der Uni- und Multivariaten Datenanalyse. *Diss*.; Fachbereich I - Psychologie, Universität Trier, Trier.